Report on BELIV'06 Workshop

BEyond time and errors: novel evaLuation methods for Information Visualization

Enrico Bertini
Dip. di Informatica e Sistemistica
University of Rome "La Sapienza"
bertini@dis.uniroma1.it

Catherine Plaisant
HCIL/UMIACS, University of Maryland
plaisant@cs.umd.edu

Giuseppe Santucci
Dip. di Informatica e Sistemistica
University of Rome "La Sapienza"
santucci@dis.uniroma1.it

DRAFT (Dec 12, 2006)

Workshop webpage:
http://www.dis.uniroma1.it/~beliv06/
Table of Contents in the ACM Digital Library:
http://portal.acm.org/toc.cfm?id=1168149&type=proceeding&coll=ACM&dl=ACM&CFID=6742925&CFTOKEN=48383460

BELIV'06 took place on May 23rd 2006, in Venice, hosted by AVI 2006 (Advanced Visual Interface Working Conference). The goal of Beliv'06 was to share experiences and new ideas about evaluation methods for information visualization. Our decision to organize the event was born from an acknowledgment of current methods' limitations, and some frustration with evaluation processes that are time consuming and too often leading to unsatisfactory results.

The purpose of information visualization is to provide users with accurate visual representations of the data and natural interaction tools to allow exploration, understanding and discovery. Success is attained when users are able to gather new non-trivial insights from the data, either finding "pearls" or making significant discoveries. These activities take place over a long time and can rarely be described with precise and predefined tasks. We lack models or methodologies to capture the process of discovery and understanding, or in fact creative activities in general. Finally, evaluation metrics such as task time completion and number of errors are judged insufficient to quantify the success of an information visualization system; thus the name of the workshop: "beyond time and errors …".

We organized a one-day workshop to gather researchers in the field to discuss evaluation issues in a informal setting. The call for papers included the following list of example topics: utility characterization and evaluation, quality criteria, metrics, insight characterization, synthetic data set generation, taxonomies of tasks, benchmark development and repositories, longitudinal case studies, adoption. We discouraged paper reporting on specific user studies unless they also included lesson learned about the methodology used. The idea was to avoid a workshop with participants merely reporting problems experienced with standard evaluation methods, but instead making progress toward newer methods and discussing their benefits and limitations over traditional methods.

The event was quite successful in attracting leading researchers in the field and gathered about 35 people, most of them active participants in informal and lively discussions. We received a surprisingly diverse set of papers covering a variety of topics. It was clear that the workshop served a unique need as many of the submitted papers addressed specific addressed in papers of traditional Infovis or HCI conferences (e.g. on the process of developing task taxonomies or synthetic datasets). We selected 15 papers, most of them proposing original and promising new ideas. Papers were made available to the participants before the workshops and authors had a chance to revise

their papers after the workshop to incorporate the comments collected during the workshop. All the papers have now been published in the ACM Digital Library and available to the larger community.

After the chairs' welcome and introduction, John Stasko summarized the topic of the workshop with a presentation called "Evaluating information evaluation: Issues and Opportunities", then each paper was given about 20 minutes (10 minutes to present and 10 minutes for discussions). We summarize the papers and discussions for each of the 4 sessions: *challenges with controlled studies*, *lessons learned from case studies*, *methodologies*, and *developing benchmarks datasets and tasks*.

Session 1 - Challenges with Controlled Studies

Controlled studies are the workhorse of evaluation: they provide solid scientific basis to formulate theories on visualization characteristics and to compare the performance of alternative solutions. Their applicability, however, is challenged by a number of practical limitations. First of all, it is very hard to design a study whose results are general enough to be used in other contexts: good selection of tasks, systems and metrics are examples of parameters hard to identify correctly. Moreover, the statistical techniques needed to draw correct figures are often difficult to handle and prone to mistakes. The two papers in this session specifically dealt with these issues. Keith Andrews reported about his own experience in using formative testing (thinking aloud), summative testing (formal experiments), and usage studies (long-term observations), highlighting for each of them specific opportunities and limitations. He argued that formative testing is useful in early development stages but the results are not useful to other infovis systems, because the adopted tasks are too narrow and specific and the number of participants is too small to have any statistical validity. Summative evaluation can be used to compare two visualizations and thus to achieve more general results; the applicability of this kind of test, however, is affected by various limiting factors. First of all, it is not easy to fairly select the visualizations involved in the comparison. In addition, it is never clear whether it is more appropriate to compare in-house implementations or the original implementations of the selected techniques. Finally, controlled studies are extremely useful to understand how users behave as they learn the system but interesting patterns of usage cannot be observed in controlled studies that are limited in time. Finally long term usage studies cannot be used to compare two or more systems. This paper and the discussions that followed highlighted how hard it is to claim success as many studies fail to report benefits, but how useful these studies are to keep developers honest about claims.

Ellis and Dix highlighted a broad spectrum of issues with controlled studies. They analysed 170 Infovis papers, showed that very few of them contained an evaluation or even mentioned evaluation as a plan for future work. The authors then make the provocative claim that the remaining papers containing some form of evaluation all had some limitations or serious problems (e.g. studies with foregone conclusions, wrong sort of experiment, fishing for results). The analysis identified complexity, diversity, and measurement as the main sources of problems in the evaluation process. Authors discussed the tension between demonstrating the goodness of a proposed technique and reporting about their potential or observed limitations. The paper raised some interesting questions about current practices and a very heated discussion followed. Some participants didn't agree with specific problems reported in the paper and warned of the danger that an overly negative review could be used by researchers who do not wish to validate their claims as an excuse to forego user evaluation altogether, a disfavour to the field. Arguments for a more balanced analysis were made.
We also discussed the problematic habit of non reporting results that do not show benefits, and the difficulty of publishing such results.

Session 2 - Lessons Learned from Case Studies

In this session we grouped the papers reporting original techniques used during the evaluation of visualizations and lessons learned from it. Henry and Fekete presented an experiment exploring

how users perceive patterns in matrix visualizations when different layout algorithms are used. The original method was to provide a printed version of the visual representation and to ask users to freely annotate their findings and mark the patterns they saw. The technique permitted to single out the effects of the different visual representations and filter out the effect of the interface. Patterns identified by all participants could be overlapped using transparency to show commonalities and differences between participants' perception and interpretations. They used tasks at different levels of detail to see how the results change when moving from basic perceptual tasks to tasks that require more complex reasoning and explored the relationship between low-level tasks such as readability and high-level tasks such as interpretation, and how they interplay in producing concept understanding.

The next two papers examined a variety of evaluation methods comparing their relative advantages. Mazza presented his experience evaluating CourseVis, a visualization system for instructors, with focus groups and compared the results with controlled experiments and semi-structured-interviews. Focus groups seemed to be particularly useful in that they permits to elicit unanticipated questions, which is not common in other types of user studies, and thus to cover a broader range of issues especially in terms of usefulness. Rester et al. compared three different solutions to aid psychologists coping with data reported by their patients. The paper comments on the relative advantages and disadvantages of their evaluation techniques: insight reports database, focus groups, and log files. Both papers agreed that a mix of methods seemed to produce better results than using a single method.

Session 3 - Methodologies: Novel Approaches and Metrics

In this session one paper described a novel methodology for long-term evaluation and two papers were about metrics-based evaluations. All of them represent new approaches. Long–term studies are, in fact, quite rare in the visualization domain, likely due to the great effort they require. Similarly, metrics-based evaluations, which can be considered part of what is commonly called model-based evaluation in the broader HCI community, are not very common in spite of their capabilities: they permit to evaluate visualizations in an objective manner.

Plaisant and Shneiderman proposed MILC (Multi-dimensional In-depth Long-term Case studies), a methodology inspired by ethnographic methods and by novel approaches found in the evaluation of creativity support tools. The paper comments on the need for long-term studies in infovis and on their potential advantages over traditional methods; notably the possibility to observe if and how users succeed in their own domain using the proposed tool. The paper gives specific guidelines on how to conduct this type of study along with tips about how to avoid common mistakes. Finally, the paper contrasts "modest" MILCs that require limited investments of resources, with more ambitious MILCs that would require significant research funding to run evaluation programs similar to clinical trials in the medical domain.

Bertini and Santucci proposed a review and analysis of "visual quality metrics", that is, a set of objective measures to quantify the *intrinsic* qualities of visualizations. The paper organized the metrics in three distinct classes: size metrics, visual effectiveness metrics, and feature preservation metrics, commenting on their scope and potential use. Emphasis was given to feature preservation metrics, a set of metrics inspired by the early Edward Tufte's work and his "lie factor" measure, but the paper instead proposes to qualify the goodness of a visualization in terms of how well it preserves the underlying data features.

The paper of Goodell et al. proposed a system to record rich session histories in a visualization environment. The tool allows for moving from the current visualization state to any of the previous ones, recording all the steps taken in the exploration. Each step is recorded by the system for further

analysis. Metrics were described that use characteristics of the graphs representing the user's paths thru the system, and the author argued that standard graph measures or specific patterns can be used to detect potential usability problems.

Session 4 - Methodologies: Heuristics for Information Visualization

After a nice Italian outside lunch provided by the AVI conference and a invigorating stroll in the narrow streets of Venice we reconvened to finish the session on methodologies. Two papers dealt with heuristics. Zuk et al. described three different sets of heuristics used to evaluate a visualization system. The results show that the kind of problems discovered by the evaluators is highly dependent on the chosen set, thus putting forward the problem of finding a minimal set of heuristics able to cover a larger group of problems. Moreover, the study highlights the potential conflicts and redundancies that can be experienced when using different heuristics sets. Discussion highlighted the benefit of having heuristics specific to Infovis, but also the challenge of having too many heuristics, making them unpractical to use.

The paper by Ardito et al. also confronted the problem of heuristic evaluation for infovis but from a different point of view. The authors proposed to adapt an established methodology called AT (for Abstract Tasks methodology), to the visualization domain. The methodology has the interesting feature of defining a precise protocol that evaluators must follow, allowing even novice evaluators to use the heuristics in a more rigorous way.

Session 5 - Developing Benchmarks datasets and tasks

This last group of papers dealt with the specific issue of providing infovis researchers with a shared pool of benchmark datasets and metrics to facilitate evaluation and allow the comparison of two or more visualization systems. Unlike other domains, where standard sets of metrics have been devised to measure the success and where a common protocol is used by researchers to assure "comparability" (e.g., Data Mining's KDD Cup http://www.acm.org/sigs/sigkdd/kddcup, or the Information Retrieval's TREC series of datasets and conferences http://trec.nist.gov), the infovis domain still lacks benchmark datasets and tasks, and commonly accepted metrics. An attempt to fill this gap is the InfoVis Contest, now hosted every year by the IEEE InfoVis Symposium. The Beliv'06 workshop hosted three papers about benchmark datasets and two about task taxonomies.

Guiard et al. proposed an experimental platform based on Shakespeare's Complete Work as a way to compare task performance in the navigation of multi-scale documents. The platform implements some basic navigation and pointing commodities (zoom & pan, perspective view, etc.) and it is meant to easily accommodate novel navigation techniques, new input devices, and a variety of contexts (Standard PCs, PDAs, tablet PCs, etc.). A researcher who devises a new technique can easily implement it in the platform comparing it with the standard approaches. This is was found a good example of how shared datasets and tasks can be very effective when designed for a small set of problems. A common debate about benchmark datasets (and tasks) is whether they should be designed to be as general as possible or tailored to a series a different areas with specific tasks and data sets.

Whiting et al. presented the "Threat Stream Data Generator" a methodology and a software tool to generate synthetic threat data. The work is part of the recent NVAC effort in visual analytics where the focus is on the utilization of analytical visualization tools to manage scenarios like disasters and terrorists attacks. First a believable scenario is designed, then the scenario is expressed in data (people, time, events, etc) and injected in an existing and large real dataset. This method allows ground truth to be known. The benefits of ground truth were discussed extensively during the workshop: having a set of findings the evaluators can anticipate permits to derive precise figures

about the effectiveness of a tool in supporting the discovery process. Furthermore, synthetic data generation is more flexible and permits to control more finely some data features (e.g. subtlety); thus allowing richer set of studies than those applicable to immutable real datasets. On the other hand, synthetic datasets might present unrealistic information and are perceived as fakes and uninteresting by analysts; which can affect the quality of an evaluation.

Melançon analysed the generation of synthetic graphs highlighting how the generation process might produce misleading results. The paper addressed a very specific problem but in great depth. It highlighted how synthetic data can be hard to produce and require special attention to generate realistic cases. One of the points of the papers, as an example, was that, in order to generate graphs similar to real world graphs (which is desirable if we want deal with real graph complexity problems and test the users behaviour in a realistic context) it is mandatory to take into account "scale-free networks" and "small world" networks.

The final paper by Lee et al. described a task taxonomy for graphs. The authors started from an established low-level task taxonomy and built with it a series of higher level objects and tasks for the specific domain of graphs. Each high level task as described with examples taken from multiple real world applications such as: FOAF (friend of a friend) networks, food webs, ontologies. The discussion focused on the challenge of refining the task taxonomies and the possible use of wikis to support this activity. As a result a section about task taxonomies was added to the infovis wiki (www.infovis-wiki.net).

Valiati et al. also proposed a task taxonomy for visualization systems. The paper comprises an extensive review of previous models and a novel taxonomy for multidimensional data. In this proposal tasks are organized in hierarchical fashion thus enabling the selection of tasks at different level of details. The authors conducted two user studies to test the taxonomy, checking for inconsistency and incompleteness. While all the observed tasks were covered by the taxonomy the authors noticed that some tasks require a finer level of description to be captured accurately and that the hierarchical organization was too rigid, since some activities may involve subtasks pertaining to different classes.

These last two papers demonstrated how task taxonomies can operate at different level of details and how useful specific sets of tasks can be for particular domains. The workshop's participants discussed the relative usefulness of low-level versus high-level tasks and specific domain versus generic tasks. Indeed, low-level tasks can not always easily be reformulated in terms of specific domains. At the same time, drawing a particular set of tasks for each domain can quickly become a daunting activity. We discussed how it is not clear how to judge the quality of one taxonomy over another. One factor seems to be the number of tasks, in order to favour compact and easily sharable taxonomies. Finally, one of the participants noted a possible misuse of the word "taxonomy". A taxonomy represents a way to organize classes of concepts so that one element cannot fall in more than one category at a time, while this is not true in the taxonomies we use.

At the end of the workshop we thanked all the participants who in turn voiced their satisfaction with the workshop format and content, and we discussed the possibility of a second workshop in conjunction with the next AVI conference in 2008.